Hereditarily Optimal Realizations: Why are they Relevant in Phylogenetic Analysis, and how does one Compute them?

Andreas Dress[1], Katharina Huber*[2], and Vincent Moulton**[3]

[1] FSPM-Strukturbildungsprozesse, University of Bielefeld, D-33501 Bielefeld, Germany
[2] Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North, New Zealand
[3] FMI, Mid Sweden University, Sundsvall, S 851-70, Sweden

Abstract. One of the main problems in phylogenetic analysis is to find good approximations of genetic distances by weighted trees. As an aid to solving this problem, it might seem tempting to consider an *optimal realization* of the metric defined by the given distances – the guiding principle being that, in case the metric is tree-like, the optimal realization obtained will necessarily be that unique weighted tree that realizes this metric. Although optimal realizations of arbitrary distances are not generally trees, but rather weighted graphs, one could still hope to obtain an informative representation of the given metric, maybe even more informative than the best approximating tree. However, optimal realizations are not only difficult to compute, they may also be non-unique. In this note we focus on one possible way out of this dilemma: *hereditarily optimal realizations*. These are essentially unique, and can also be described in an explicit way. We define hereditarily optimal realizations, discuss some of there properties, and we indicate in particular why, due to recent results on the so-called *T-construction* of a metric space, it is a staight-forward task to compute these realizations for a large class of phylogentically relevant metrics.

1 Optimal Realizations

Given a metric (or, more generally, a distance function) d defined on a finite set X of taxa representing, say, their genetic distance, one of the main problems in phylogenetic analysis is to find a *weighted tree* $T = (V, E, w)$, with V a finite set of vertices containing X, and $w : E \to \mathbb{R}_{\geq 0}$ a weighting of the edge set E of T, so that the distance d_T on X, defined by taking the weight of the shortest path between pairs of elements in X considered as vertices of T, approximates d "optimally". Many methods have been introduced to deal with this problem, and most of them satisfy the following criterion: If there is some (necessarily unique) weighted tree T with $d = d_T$, then the method returns T. As is well known [6,23], those metrics d that can actually

* The author thanks the New Zealand Marsden Fund for its support.
** The author thanks the Swedish National Research Council (NFR) for its support (grant# M12342-300).

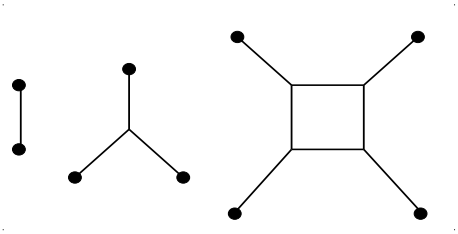

Fig. 1. By labeling the vertices and weighting the edges of these graphs appropriately (giving the same weight to parallel edges), one can obtain a unique optimal realization for any metric defined on two, three, or four points.

be represented by a weighted tree – often called *tree-like* metrics – can be characterized as those satisfying the *four-point* condition:

- For all $x, y, u, v \in X$,

$$d(x,y) + d(u,v) \leq \max\{d(x,u) + d(y,v), d(x,v) + d(y,u)\}.$$

What has also been shown [21] is that the weighted tree T that represents such a metric d is actually an *optimal realization* of d, i.e. defining the *total weight* of a weighted graph $G = (V, E, w)$ to be $\|G\| = \Sigma_{e \in E} w(e)$, and saying that G *realizes* a metric d on X if $X \subseteq V$ and $d = d_G$ (the distance on X induced by using shortest paths in G) holds, the tree T is a graph that realizes d and that has minimal total weight amongst all such graphs.

Hence, since – in practice – a given metric d does not usually satisfy the (highly non-generic) four-point condition in practice, it seems reasonable to ask for an optimal realization of d in any case. For, even though we would not necessarily obtain a tree, we would at least obtain a network realizing d that might give us a better understanding of the properties of X encoded by d. However, such a strategy has two major drawbacks: Even though, for any given metric d, an optimal realization of d is known to always exist, and even though we can obtain a unique optimal realization for d by appropriately labeling and weighting one of the graphs in Figure 1 for any X with $\#X \leq 4$, optimal realizations are hard to compute [1] and, perhaps more importantly, an optimal realization of d is not necessarily unique if $\#X \geq 5$.

In Figure 2 for example, we present two distinct optimal realizations for a metric defined on a five element set, an example that originally appeared in [9, (A 3.3)]. Indeed, it has even been shown in [1] that there exist finite metric spaces with a *continuum* of optimal realizations – see Figure 3 for an example that originally appeared in [1].

	a	b	c	d	e
a	0	7	4	5	3
b	7	0	5	6	4
c	4	5	0	7	3
d	5	6	7	0	4
e	3	4	3	4	0

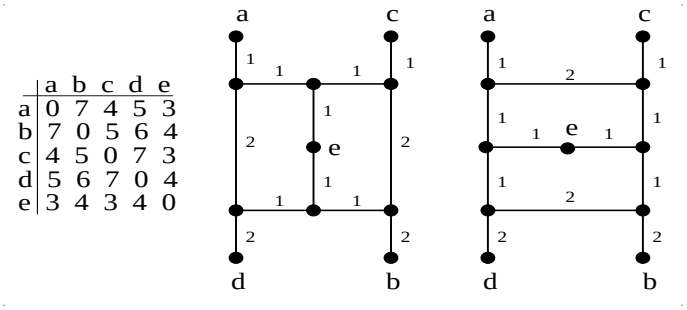

Fig. 2. A distance matrix with two distinct optimal realizations.

	a	b	c	d	e	f
a	0	1	2	2	3	2
b	1	0	1	3	4	3
c	2	1	0	2	3	3
d	2	3	2	0	1	2
e	3	4	3	1	0	1
f	2	3	3	2	1	0

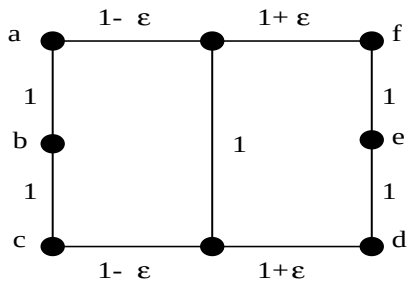

Fig. 3. For each ϵ with $\frac{-1}{2} \leq \epsilon \leq \frac{1}{2}$ the weighted graph on the right is an optimal realization of the metric on the left.

2 Hereditarily Optimal Realizations

Taking this into account, we now focus on one possible way out of this dilemma that also originally appeared in [9]: We define a *hereditarily optimal* – or, for short, an *h-optimal* – realization of d inductively with respect to $\#X$: If $\#X \leq 2$, then any optimal realization of d is defined to be h-optimal. If $\#X = k$ and h-optimal realizations have been defined previously for any metric defined on a set Y with $\#Y < k$, then a realization $G = (V, E, w)$ of d is defined to be h-optimal if $\|G\|$ is minimal with respect to the property that, for any $Y \subset X$, there is some $E' \subseteq E$ with $Y \subseteq V' := \cup E' := \cup_{\{u,v\} \in E'} \{u, v\}$ such that $G' := (V', E', w|_{E'})$ is an h-optimal realization of $d|_Y$.

Even though, at first sight, it might appear that finding h-optimal realizations could be even harder than finding optimal ones, and that they are even less likely to be unique, the converse – actually – is true: Consider the set $P(X, d)$ consisting of those functions $f : X \to \mathbb{R}$ that satisfy the condition

$$f(x) + f(y) \geq d(x, y)$$

for all $x, y \in X$. To each $f \in P(X,d)$ associate a graph $K(f)$ that has vertex set X, and whose edge set consists of those subsets $\{x,y\} \subseteq X$ for which

$$f(x) + f(y) = d(x,y)$$

holds. Next, define a weighted graph $\Gamma_d = (V_d, E_d, w_d)$ with vertex set

$$V_d := \{f \in P(X,d) \,:\, K(f) \text{ is connected and not bipartite}\},$$

edge set

$$E_d := \{\{f,g\} \subseteq \binom{V_d}{2} \,:\, K((f+g)/2) \text{ is connected and bipartite }\},$$

and weighting

$$w_d : E_d \to \mathbb{R}_{>0}$$

defined by

$$w_d(\{f,g\}) := \sup_{x \in X} |f(x) - g(x)|.$$

Note that we can consider X as being a subset of V_d – simply associate the map $h_x : X \to \mathbb{R}$ defined by $h_x(y) := d(x,y)$ for all $y \in X$, to each element $x \in X$.

Then, it follows from [9, Theorem 7] that V_d is finite, and that the following hold:

- The graph Γ_d is an h-optimal realization of d.
- If $\Gamma = (V, E, w)$ is any other h-optimal realization of d, then Γ is essentially isomorphic to Γ_d i.e. it becomes isomorphic to Γ_d once vertices $v \in V - X$ of degree two have been deleted one by one and the corresponding edges $e_1 = \{u_1, v\}, e_2 = \{v, u_2\} \in E$ have been replaced by the single edge $\{u_1, u_2\}$ that is given weight $w(\{u_1, v\}) + w(\{v, u_2\})$.

Thus, h-optimal realizations have the advantage that they can not only be described in an explicit way, but they are essentially unique, too. Moreover, it follows from results contained in [9] that Γ_d is isomorphic to the unique weighted tree representing d whenever d is tree-like.

3 A connection with T-theory

Although the definition of Γ_d might seem a bit strange, within *T-theory* it makes a lot of sense [19]. In fact, considering the set $P(X,d)$ as an unbounded polytope in $\mathbb{R}^X$, it is shown in [9] that Γ_d is a subgraph of the weighted graph $T_d^{(1)} = (F_0, F_1, w_P)$ that has vertex set F_0 consisting of the 0-dimensional faces (i.e. vertices) of $P(X,d)$, edge set F_1 consisting of those $\{f,g\} \in \binom{F_0}{2}$ for which f and g are the vertices of a 1-dimensional face in $P(X,d)$, and weighting w_P defined in exactly the same way as w_d was defined above. The

graph $T_d^{(1)}$ is better known as the 1-skeleton of the *tight span* $T(X,d)$ of d, a polytope consisting of the compact faces of $P(X,d)$, and the main object of study in T-theory. Thus, in theory at least, it is possible to compute $T_d^{(1)}$, and hence Γ_d, using any of the many packages that are available for computing polytopes.

We now consider a simple example of an h-optimal realization: As a consequence of [21, Theorem 3.2], the complete bipartite graph $G := K_{3,3}$ (in which every edge is assigned weight one) is an optimal realization of the metric d_G induced by G on its vertex set. Moreover, the h-optimal realization of d_G also coincides with G, but it can be seen that $\Gamma_{d_G} \neq T_{d_G}$ holds (see Figure 4). Hence, it may be of some interest to characterize those metrics d for which $\Gamma_d = T_d^{(1)}$ holds.

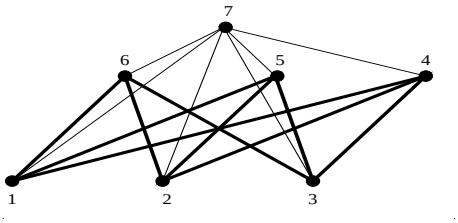

Fig. 4. The graph $T^{(1)}_{d_{K_{3,3}}}$ (all edges have weight one) with the subgraph $\Gamma_{d_{K_{3,3}}}$ $(= K_{3,3})$ indicated in bold.

In [14], we give an explicit answer to this question: Defining, for any four elements $u,v,x,y \in X$,

$$d(xy|uv) := \max\{d(x,u)+d(y,v), d(x,v)+d(y,u)\} - d(x,y) - d(u,v)$$

and putting $\alpha(xy|uv) := \max(d(xy|uv),0)$, we see that those metrics d for which $\Gamma_d = T_d^{(1)}$ holds can be characterized by the following *five-point* condition:

- For all $t,x,y,u,v \in X$, the inequality

$$d(xy|uv) \leq \alpha(xt|uv) + \alpha(xy|ut)$$

holds.

Such metrics are called *totally decomposable* [2]. As a consequence of recent results appearing in [11–13,15], we also see in [14] that if, in addition, d satisfies the following *six-point* condition

- For every subset Y of X of cardinality 6, there exists a pair $u, v \in Y$ of distinct elements with

$$0 \leq d(xy|uv)$$

for all $x, y \in Y - \{u, v\}$ (and hence for all $x, y \in Y$),

then the underlying graph (V_d, E_d) of $\Gamma_d = (V_d, E_d, w_d)$ is isomorphic to the well-known *Buneman graph* [10], a graph that has been used previously with quite some success in phylogenetic analysis [4,5,16]. Thus, in summary we have the following result for metrics that satisfy the above five- and six-point conditions, metrics that are called *consistent* for short [15]:

Theorem 1. *If d is a consistent metric, then the weighted Buneman graph Γ_d is an h-optimal realization of d.*

4 Concluding Remarks

Finally, these results give us a framework for better understanding the uses of the *split-decomposition* technique in phylogenetic analysis [2,3]: Define d to be a *Kalmanson* metric [7,8,22] on X if there exists a labeling $X = \{x_1, x_2, \ldots, x_n\}$ $(n := \#X)$ of X such that $1 \leq i < k < j < l \leq n$ implies

$$d(x_i, x_j) + d(x_k, x_l) \geq \max(d(x_i, x_k) + d(x_j, x_l), d(x_i, x_k) + d(x_j, x_j)),$$

and note that the following inclusions for the various types of metrics that we have considered above hold:

$$\text{tree-like} \subseteq \text{Kalmanson} \subseteq \text{consistent} \subseteq \text{totally decomposable}.$$

Moreover, if d is a Kalmanson metric, then it can be realized by a certain *outer-planar* graph [17] that is an isometric subgraph of Γ_d and is used by the program *SplitsTree* [18,20].

Further, even though a given metric d will not in general be totally decomposable, split-decomposition theory provides us with a unique maximal totally decomposable submetric d_{split} of d that satisfies the condition $P(X, d) = P(X, d_{split}) + P(X, d - d_{split})$ [2].

In practice, d_{split} tends to be Kalmanson so that it can be represented by the outer planar graph mentioned above. If d_{split} is not Kalmanson, then some variant of the Buneman graph is currently used by SplitsTree. In light of the above results, however, the graph $\Gamma_{d_{split}}$ (or some isometric subgraph of it) might be more appropriate in this situation. Thus, in conclusion, it appears worthwhile to develop new techniques for efficiently computing the h-optimal realization of d_{split} or even that of d, a task that – as we have seen above – is closely related to computing their tight spans [13].

References

1. Althöfer, I.: On optimal realizations of finite metric spaces by graphs, Discrete Comput. Geometry **3** (1988) 103–122
2. Bandelt, H.-J., Dress, A.: A canonical decomposition theory for metrics on a finite set, Adv. in Math. **92** (1992) 47–105
3. Bandelt H.-J., Dress, A.: Split decomposition: a new and useful approach to phylogenetic analysis of distance data, Molecular Phylogenetics and Evolution **1 (3)** (1992b) 242–252
4. Bandelt, H.-J., Forster, P., Sykes, B., Richards, M.: Mitochondrial portraits of human population using median networks, Genetics **141** (October 1995) 743–753
5. Barthélemy, J., Guenoche A.: Trees and Proximity Representations, John Wiley & Sons, Chichester New York Brisbane Toronto Singapore, 1991
6. Buneman, P.: The recovery of trees from measures of dissimilarity, In F. Hodson et al., Mathematics in the Archeological and Historical Sciences, (pp.387-395), Edinburgh University Press, 1971
7. Chepoi, V., Fichet, B.: A note on circular decomposable metrics, preprint (1998)
8. Christopher, G., Farach, M., Trick, M.: The structure of circular decomposable metrics, ESA 1996
9. Dress, A.: Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: A note on combinatorial properties of metric spaces, Adv. in Math. **53** (1984) 321–402
10. Dress, A., Hendy, M., Huber, K., Moulton, V.: On the number of vertices and edges in the Buneman Graph, Ann. Combin. **1** (1997) 329–337
11. Dress, A., Huber, K., Moulton, V.: Some variations on a theme by Buneman, Ann. Combin. **1** (1997) 339–352
12. Dress, A., Huber, K.T., Moulton, V.: A Comparison between two distinct continuous models in projective cluster theory: The median and the tight-span construction, Ann. Combin. **2** (1998) 299-311
13. Dress, A., Huber, K.T., Moulton, V.: An explicit computation of the injective hull of certain finite metric spaces in terms of their associated Buneman complex, preprint (1999)
14. Dress, A., Huber, K.T., Moulton, V.: Hereditarily optimal realizations of consistent metrics, in preparation
15. Dress, A., Huber, K.T., Koolen, J., Moulton, V.: Six points suffice: How to check for metric consistency, in preparation
16. Dress, A., Huber, K.T., Lockhart, P., Moulton, V.: Lite Buneman networks: A technique for studying plant speciation, preprint (1999)
17. Dress, A., Huson, D.: Computing phylogenetic networks from split systems, preprint (1999)
18. Dress, A., Huson, D., Moulton, V.: Analyzing and visualizing distance data using SplitsTree, Discrete Applied Mathematics **71** (1996) 95–110
19. Dress, A., Moulton, V., Terhalle, W.: T-theory: An Overview, Europ. J. Combinatorics **17** (1996) 161–175
20. Huson, D.: SplitsTree: a program for analyzing and visualizing evolutionary data, Bioinformatics **14 (1)** (1998) 68–73
21. Imrich, W., Simoes-Pereira, J., Zamfirescu, C.: On optimal emdeddings of metrics in graphs, Journal of Combinatorial Theory, Series B, **36, No.1**, (1984) 1–15

22. Kalmanson K.: Edgeconvex circuits and the travelling salesman problem, Canadian Jour. Math., **27** (1975) 1000–1010
23. Zaretsky, K.: Reconstruction of a tree from the distances between its pendant vertices, Uspekhi Math. Nauk (Russian Mathematical Surveys) **20** (1965) 90–92 (in Russian)